CONSTRUCTION, COLLECTION & CURATION OF NASA'S DATA REFERENCE MODELS

NAVIGATING NASA'S INFORMATION SPACE

Inaugural Report from NASA's Enterprise Architecture Data Team

August 1st 2006

Abstract:

In February 2006, the Enterprise Architects of NASA identified the need to develop a strategy that would help NASA be more consistent about our use of, reliance on, and trust in our data, and which would enable information sharing and reuse. This paper describes a practical strategy for organizing our information and data assets so they can be discoverable (by machines and humans) and reusable. Within this paper we describe the nature of NASA's information problem and recommend specific design goals and design principles that should be employed to solve it. Additionally, we make specific technical recommendations including the development of a curation-friendly catalog of NASA applications, service level agreements, access policy agreements, and data reference models. We recommend a strategy for the Enterprise Architects to join with the larger community of practitioners within NASA and combine our efforts to greater effect and we list and describe current NASA data models as notable examples from this community. Finally, we list a series of activities and a proposed timeline. These initial recommendations, if adopted and supported, will have a positive impact by establishing conventions, standards, and best practices for information to be more easily shared, trusted and reused. It will have a positive effect on establishing a data re-use culture; it will have an impact on our concepts of data sourcing, and provide us a clearer understanding of what information and models have been validated and what can be trusted.

Contents:

Background:

Our reliance on data and the information we derive from it touches everything that we do. In February 2006, the Enterprise Architects of NASA identified the need for a strategy that would help us to be more consistent about our use of, reliance on, and trust in our data, and which would enable information sharing and reuse. Our goal is to describe a practical strategy for organizing our information and data assets so they can be discoverable (by machines and humans) and reusable. We also recommend a strategy for the Enterprise Architects to join with a larger community of practitioners and combine our efforts to greater effect. These initial recommendations, if adopted, will have a positive impact by establishing conventions, standards, and best practices for information to be more easily shared, trusted and reused. It will have a positive effect on establishing a data re-use culture; it will have an impact on our concepts of data sourcing, and provide us a clearer understanding of what information and models have been validated and what can be trusted.

Problem:

Critical information related to our daily operation is becoming increasingly more difficult to find. It is difficult to find relevant information that you *know* is available and virtually impossible to discover critical information that is relevant but unknown. When we cannot find resources, we often recreate them. When we have trouble integrating information, we often copy it. These habits make NASA's data volume and data integrity problems worse.

For decision-making support, it is impractical to pre-determine the complete set of applicable knowledge or every possible query required to satisfy each individual's needs. Use cases and requirements change all the time. We cannot anticipate in advance what the next collection of information elements must be, or for what purpose. Flexible queries and effective access (as opposed to hard-coded responses) are what is needed.

NASA's information is voluminous, diverse, and extensive, has impacts on our entire community, and is growing at an unabated rate. And our information problem exists within at least five dimensions: size, complexity, diversity, rate of growth and trust. NASA's compliance with the Federal Data Reference Model serves as an additional incentive for us to look inward and find practical, cost effective solutions to facilitate information discovery and reuse.

The Customer Experience:

How can we make information easier to find and, once it is discovered, make it easier for the next person to discover? How can our experience using the information benefit the next person that needs it? How can NASA leverage our network and web infrastructure, and our data management disciplines? How can we make information contained within databases and systems across projects and programs discoverable without disruption, without great expense, without loss of original contextual meaning, and without breaches of trust?

We propose a mechanism for customers to easily find information services. Essentially, a database of databases, but unlike other databases, this one is built from the fabric of the Web. Based on policy rules, and by using current web standards and technologies, computers will be able to negotiate with each other for access and services. Also, because our proposed design is based on web technology, customers will be able to browse, query, and search through NASA's collection of information resources as easily as choosing a hotel or sweater. Eventually, use pattern matches will assist customers. Similar queries and browsing experiences from others will be made available to you. You will be able to determine, through your browser, attributes of the information's currency, provenance, validity, and trust. Further, like the Web it is part of, this information service's utility will be enriched by each customer's use over time and it will grow incrementally, just like the Web.

EA Data Team's Goals for Information Management:

- Establish Information Management standards and mechanisms that promote enriched and ad-hoc information sharing and reuse across NASA data services.
- Define a prospective solution that will augment data management capabilities as newly created data sources are integrated.
- Integrate in a layered approach, enriching services incrementally, when practical and requirements-driven.
- Enable integration so that the most sought after, useful, and mostly easily integrated data services (databases, models, web services, etc.) are pushed to the front of the queue.
- Enable discovery and reuse of policy agreements between data providers and customers and across data systems so attributes of confidentiality, integrity, availability and currency are managed uniformly across diverse systems.
- Enable easier query integration across disparate hierarchies by modernizing NASA Information Standards to include a NASA Data Reference Model and definition of "gold, silver and bronze" standards for data and data models.
- Leverage current communities who have demonstrated excellence within their projects and programs.

EA Data Team's Design Principles for Information Management:

Information is Strategic – Our reliance on and the importance of information is a basic, underlying fact of all NASA activities. Effective stewardship is, therefore, critical. This includes structured data, semi-structured data, unstructured data, databases, instrument data, reference models, drawings, text, and photographs.

Open Internet Standards – The diversity of our systems requires adherence to standards that were designed specifically to work in a highly diverse and distributed environment. Strict conformance to these standards enables communication and interoperability across widely disparate systems. Proprietary variants of these standards are to be avoided.

Loose Coupling – Avoids stovepipes by providing a resilient relationship between two or more computer systems that are exchanging data by enabling each end of the transaction to make its requirements explicit and make few assumptions about the other end. Loosely Coupled systems are considered useful when either the source or the destination computer systems are subject to frequent changes. *

Machine-Readable - Whenever practical, information integration infrastructure must conform to machine-readable standards, thus minimizing the expense of customized or human interfaces.

Make NASA Information Models Discoverable and Reusable – Discovery and sufficient information about data models enable customers (machines and humans) to determine suitability for reuse.

Formalization of essential attributes – Knowledge of how or if an information service has been validated at both the data or logic level, what its currency values are, who its author is, and, most importantly, if it functions as intended, needs to be uniformly understood across diverse systems.

Federal versus NASA Data Reference Model:

The intent of the Federal Data Reference Model (DRM) is to promote data sharing and reuse across Federal Agencies. Given the diversity of Federal agencies it is understandable that the Federal DRM is non-specific regarding standards for organizing data in a description format. The current DRM is meant as a starting point and virtually any data model standard will fit. However, too much flexibility may inhibit tangible progress toward NASA's requirements for information discovery, reuse and trust. We recommend leveraging both the intent of and the deliverables to the Federal DRM by formalizing NASA's collection of NASA Data Reference Models.

The EA Data Team recommends establishing such mechanisms that will enable our systems to find, understand, negotiate, and provide information that is contextually relevant, but which will not impact our current data sources by introducing unreasonable code or format changes. We believe that a formalized mechanism to curate these DRMs without negative impact to specialized data classifications is essential for our data management strategy. There is a wide range of complexity, utility and purpose to many data models being successfully used within NASA. We need to maintain their contextual relevance while making them more accessible and usable.

An Information Library Service provides an infrastructure for Data Integration Policies and Service Level Agreements

Because of NASA's unique data challenge, the need for information integration development projects is especially acute and essential. Given the heterogeneity and diversity of NASA data (e.g., scientific, administrative, operational, financial, analytic), we need a flexible approach to building information integration solutions with sufficient formality to provide cross-system discovery and reuse. The most successful distributed information system ever conceived, the World Wide Web, offers plenty of inspiration in areas such as web services, Web 2.0, and Semantic Web.

As NASA embarks on wide-scale information integration projects, an information management problem soon becomes apparent: how do we manage all the disparate data sources so that duplication, inflexibility, and errors are effectively managed?

Our response to this problem is to create a new declarative data format that is machine-readable, and then push the management problem onto the computer and the infrastructure. We propose to do the same thing for information integration and enterprise architecture problems by creating a Semantic Web data representation for Service Level and Access Policy (SLAP) Agreements.

Using W3C Semantic Web standards (WSDL, OWL, RDF, SWRL and SPARQL), we recommend creating a machine-readable data format to represent information about data sources and information integration applications.

How to access the data:

- Web Service Description Language (WSDL)
    - Describes the inputs and outputs of applications (including data sources applications).
    - Describes the network service endpoints where the data may be accessed
    - Describes the details of access mechanisms (HTTP, SOAP, JMS, etc)

All the various, disparate *policies* surrounding information integration:
Our recommendation for building high-quality information integration projects is *policy management*; this refers to the details of the service-level agreements that exist between data producers and data consumers. Given the scale and complexity of NASA's challenges, the policy requirements (and the sheer scope of the number of policies) make the problem solvable only by means of computer management. *Some* of the policy questions include:

- Who can access the data?
- When can they access it? (How often, over what duration, etc)
- Why would someone want to access this data, i.e., what is the data good for?
- Where does the data originate from?
- What curation processes control the data quality?
- What is the carrying capacity of the application that supports this data source?
- Is there spare capacity for accessing the data source?
- What can clients grant/do with the data? What can they not do?

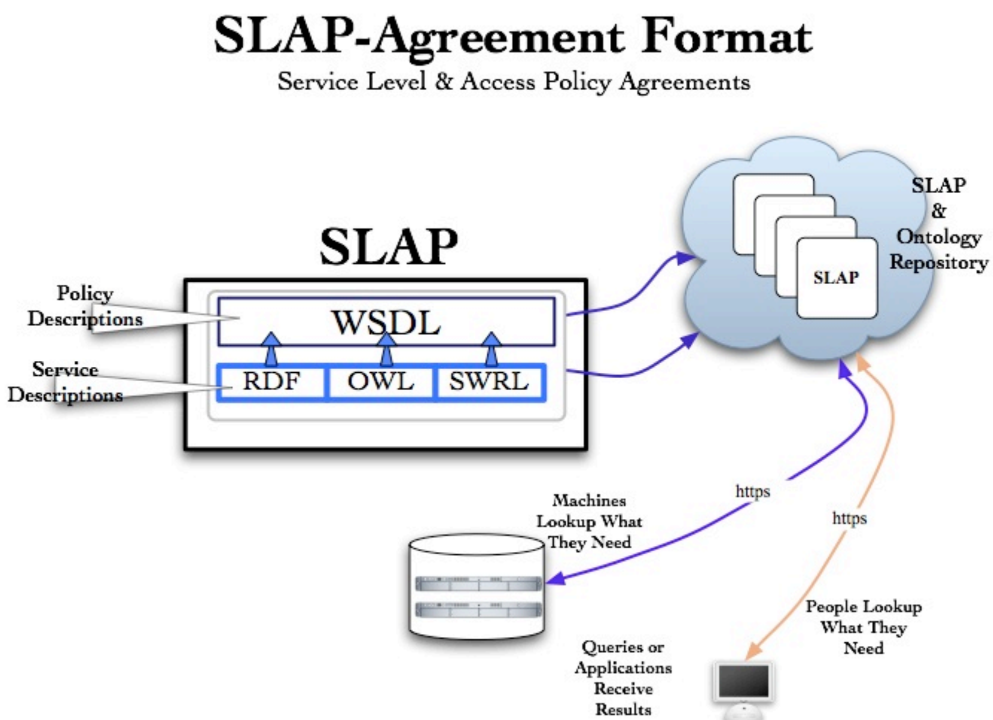
These policy languages are, for the most part, NASA-specific. While other similar organizations face similar problems, policies are typically embedded in and embodied by particular programs, projects, or situations. Thus, while it is important to use common data representation standards to create machine-readable policy languages for NASA, the job of creating those languages must be performed either by NASA personnel or by those who are familiar with NASA's institutional challenges and cultures -- most likely by some combination of both groups.

We propose, therefore, to use the Semantic Web standards RDF and OWL, as well as SWRL and other rule formalisms, to represent NASA information integration policy languages in semantically rich way. Such policy languages have the following benefits:

- They support accurate, declarative representations of policies. The languages are expressive enough to permit the writing of policies that closely capture the full intent of the policy writer in a fairly natural way.
- There are a variety of powerful analytical services for these languages. For example, one can have an automated reasoner check for policy containment, redundancy, or incompatibility. Such services are critical for the effective management of large numbers of changing policies.
- There is a good infrastructure for these languages including editors, reasoners, debuggers, and visualizers.
- There are a growing number of domain ontologies expressed in these languages (see SWEET p.13). Since information policies are centered on the

> subject matter of the information they control, it is essential to have sufficiently well modeled theories of that subject matter. These languages give us seamless access to such.

SLAP Agreements provide a necessary process for providing and gaining access to information resources. Additionally, since the information contained in the SLAPs is searchable, they provide a useful reference for other SLAPs reflecting data in use, not merely data that is available. Most importantly, discovered information is left for others to find, enrich, reuse, and leave for others to do the same. Each customer enriches information elements about services or models with each use.



Using the latest work on networking grids and electronic markets, it will be possible to add a robust set of services around negotiation of SLAs to facilitate information integration projects and to solve resource contention and allocation problems. Bargaining between individual participants and candidate participants in SLAPs solves these issues.

Policy Awareness and Flexibility at the Data Level

By constructing the SLAP agreements in conformance to a generalized ontology we will be able to adjust specific policies based on many different situational conditions. Design and modeling of the SLAP ontology should be a near-term activity. For conceptual reference, the SLAP model might be constructed with the following concepts.

hasCapacity (capacity = the number of parties [int] that can use the service)
hasParty (party = who is a party to the SLA, a person, a manager, an organization)
hasPolicies (a discussion to determine if WS-policy is sufficient is required *potential* policy examples include: Who can access, When can they access, How often over a period of time can they access, Criteria on data release based on multiple conditions)
hasDataSource (data source = data or DRM feed)

DataSource

- hasDomain (data source domain = type)
- hasServiceDescription (WSDL or might have a URI to a graph)
- hasProvenance (provenance = who created is responsible)
- hasValidity (validity = the formal validation)
- hasCurrancy(currency = date of creation and or last update)
- hasTrust (trust = reliability or accrued instances of success)

Information Integration Policy Requirements:

Human decisions are often based on data or information provided by a computer. In order to quantify, or even understand the soundness of a decision, certain fundamental facts about the data are required. Many people assert that their data is credible by declaring that it came from an "authoritative source", but what they often mean is that the data originates from an organizationally responsible source. While knowledge of the source (provenance) is valuable, it is not necessarily equivalent to authoritativeness or credibility. To determine whether data is credible, knowledge of validity, currency and trust must also be understood. SLAP documents will provide a mechanism to discover these four essential elements. An understanding of who or what generated the data, when it was generated, whether it has been validated, and if it is being used successfully will improve the confidence of decisions that rely on it and provide additional opportunities for data discovery and reuse.

Provenance. Subject to access controls, anyone (and any computer) should be able to discover who or what created a database and who or what generated data. Aside from a basic "none-available" or a name, attributes of provenance should include organizational or project context. For example, "Created by Space Operations Mission Directorate, Shuttle Avionics Guidelines, Compiled by A.A. Aaaaaa/NASA, Approved by B.B. Bbbb for the Strategic Avionics Technology Working Group." Levels of granularity need to be identified but originator of the data is essential.

Validity. Subject to access controls, anyone (and any computer) should be able to discover whether or not the data or the database has been validated and by what mean it has been validated. Attributes of validity should include information regarding various validation mechanisms. For example, indications that an information model was processed by a reasoner, and that the logic of the model proved not to contain any flaws. If there were an assertion that a Commander of a Space Shuttle must be a US citizen, but does not need to be a NASA employee, it would be appropriate to cite the Human Resources regulation that stipulates these conditions.

Currency. Subject to access controls, this allows a person (or a machine) to determine when the data was last updated. Possible variants of currency include: original date content was created; date first uploaded into system; date the content was published publicly; date of the last modification to the content; date source was last used by another system; date content was removed from the system.

Trust. Unlike explicit instances of date, authorship, or proofs, a trust attribute enables any person (or any machine) to determine how often or whether a data service has been used successfully. By showing how many parties participate in a data sharing agreement, SLAP-Agreements would show how often a service has been used. While not the sum total of trust, it can indicate how often a wide population of customers uses a data element or collection successfully.

Ontology Curation

Ontologies (and other kinds of data reference model technology) are intended to be shared models; they are, after all, mechanisms by which humans and computers come to understand a knowledge domain formally and rigorously. The intent and scope of shared models range from highly curated, large scale, validated systems to casual, ad hoc experiments. The use of models (of whatever level of development) requires adaptation, evolution, and communication. If the models are effectively shared, then they must be living documents, and the processes – both institutional and technical – for sharing them should encourage their appropriate growth. Reusing ontologies by copy-and-paste-then-modify has all the problems of cut and paste programming code reuse: it is fragile, time consuming, and makes it difficult for improvements to spread to all users of the ontologies from arbitrary other parties.

To support the interactive and machine-mediated sharing of ontologies, we need an Agency-wide ontology repository system. Such a system would consist of

- Publishing mechanisms, both for the ontology documents and the metadata about them
- A collection of Web Services that provide a variety of search, comparison, analysis, and validation services
- Integration with the ontology creation tools
- Extensive communication and feedback mechanisms

In short, we need a curation mechanism, but also a way to enable growth. We need centralization, but freedom at the edges too. Without non-coordinated growth in separate directions, NASA won't evolve the ontologies, DRMs or policies it needs. But without some means to resolve the tensions inherent in such growth, NASA won't have the shared means of expression that it needs. The Agency-wide ontology and reference model repository should be designed with this tension in mind.

Examples: Start By Leveraging Our Current Data Reference Models

As noted earlier, because the Federal DRM definitions are so broad, many NASA examples could be used should we be called on to produce an inventory. However internally, NASA has a community of experts and a culture of rigorous data management that positions us favorably to set standards that leverage our own experience and requirements to take the next step toward information integration. Our Scientific, Research and Mission Operations communities, who have a major interest in getting things right. We have many well-curated collections ranging in complexity and utility, so while this is not an exhaustive listing, the following examples reflect practical data models in our community today.

BIANCA – Headquarters Information Technology and Communications Division
The BIANCA (Business Impact Analysis for Networked Computer Assets) application is a database integration project for network assets at HQ. It integrates information about networked computer assets and allows users to browse the properties of these entities, as well as their interconnections, including connections between applications (sources and sinks), servers, networks, and network services. BIANCA analyzes dependencies between these assets in order to provide services like repair plans, outage cost estimates, and dependency reports. Users can query across the federated information store and browse the data in a web browser in order, for example, to track the impact that a failure of one system, subsystem, or application would have on other systems and customers. The BIANCA RDFS data reference model can be reused by other applications including real-time collection of configuration data or change processes (configuration changes over time).

Business Information Warehouse – The Integrated Enterprise Management Office
The main data store for information delivery for IEMP is in SAP's Business Information Warehouse (BW). BW consists of validated information in several different models including the relational persistent staging area and operational data store, and multi-dimensional cubes based on extended star schemas. The BW models create a standardized view of agency financial information based on the standard federal accounting events of Commitments, Obligations, Costs, and Disbursements (COCD), labor information based on the Agency Labor Distribution System, travel information based on the Travel Manager system, and procurement information based on the Contract Management Module (to be released). The model comprises data that results from all agency transactions, including budget execution, purchases, reimbursable orders, travel, labor, and procurement. COCD information can be analyzed from multiple dimensions, including Mission, Theme, Program, Project,

Business Area (NASA Centers), Fund, Fund Center, Cost Center, Fiscal Year, and Programmatic Year. This information is used by many people throughout the Agency to perform their jobs, including: Resource Analysts and Managers in support of Program and Project managers to review costs and available resources, Budget Analysts to provide insights into available budget to be used by their project, Cost Analysts to monitor cost of work performed, and the Office of Chief Financial Officer to analyze the Agency's financial status.

GENESIS SciFlo: Multi-Instrument Climate Science Using Grid Workflow (a REASoN project at the Jet Propulsion Laboratory)

NASA's Earth Observing System (EOS) is the world's most ambitious facility for studying global climate change. The mandate now is to combine measurements from the instruments on the three flagship platforms—Terra, Aqua, and Aura—and other Earth probes to enable large-scale studies of climate change over periods of years to decades. To enable large-scale, multi-instrument atmospheric science, the REASoN-funded GENESIS (General Environmental & Earth Science Information System) project at JPL has developed the SciFlo (Science Dataflow) workflow engine. A distributed network of SciFlo execution nodes is being deployed this year at several universities and several NASA Earth Science centers, including the Distributed Active Archive Centers (DAACs) at JPL, Langley, and Goddard. The SciFlo network enables researchers to tie together local analysis algorithms and remote Web Services into a distributed dataflow, inject custom data selection, mining and fusion operators and services into the DAACs, and thereby efficiently generate *custom, multi-instrument* products for *large-scale* science investigations. There is a happy synergy between semantics and structured workflow: workflow engines can *use* semantic metadata for logical inference; conversely, a structured workflow system provides an opportunity to *capture* and *preserve* semantic metadata and data lineage, and *infer* additional semantic annotations. Smart workflow systems that choreograph services will benefit NASA in many ways by enabling users to: publish, discover and reuse versioned algorithms as services; rigorously specify reusable analysis flows, publish flows and exchange them with colleagues; implement new composite services by authoring a workflow; query the provenance of generated products; trace the effects of data or processing anomalies; modify & repeat large-scale science analyses, and more.

Integrated Collaboration Environment – Office of Exploration

The NASA Exploration Information Ontology Model (NExIOM) is a proposed common data model for use within the Exploration Systems Mission Directorate (ESMD). The scope of the model is quite broad, encompassing all ESMD project phases from requirements analysis and concept formulation through design, manufacture, training, and operations. NExIOM is the proposed method through which ESMD tries to reconcile all of the information involved in these activities: what the data means, where it is located, and how it can be applied. NExIOM would support defensible decision making through consistent, traceable, and understandable data representation. To accomplish these goals NExIOM will provide a standards-based, common language definition for engineering terms used to represent ESMD product architectures (vehicles, missions, and technologies) and development architectures (tools, models, simulations, processes, decisions, requirements), as well as meta-data about tool applicability, lineage, and accreditation. The NExIOM data model includes an ontology, a set of schemas, and a data dictionary. NExIOM data would be stored in the ESMD Integrated Collaborative Environment (ICE).

Meta Data Management System – Office of the Chief Financial Officer

The Metadata Manager (MdM) is a web-based system that manages the Agency's official NASA Structure Management (NSM) data elements and associated codes. The MdM project, owned by the Office of the Chief Financial Officer (OCFO), established a methodology and workflow for aligning the Agency's technical WBS with NASA's financial coding structure. MdM is a workflow tool used for identifying, creating, approving, tracking, organizing, and archiving the Agency's structural codes for: Appropriation, Mission, Theme, Program, Project and WBS 2 through WBS 7 structural elements. NASA Program Managers and other "code requesters" create and maintain these codes, which are used for the N2 Budget Formulation System, Core Financial IEM/SAP System, and project management systems. The NSM coding structure, maintained by MdM, provides NASA with the capability to satisfy the Agency goal to manage Programs using Earned Value Management (EVM).

POPS – Office of the Chief Engineer

The POPS Project (People, Organizations, Projects, and Skills) reference model was constructed in OWL and RDF Schema to represent the class and property relationships of several data sources about NASA employees, their competencies, and their project assignments. It was constructed as a mechanism for information integration and is used to query across data sources without modification to the original sources. And because of the formal, model-theoretic semantics of RDF and OWL, the integration model is known to be formally consistent and coherent. That is, using RDF Schema and OWL means that the integration model is machine-readable and provides consistency checking; classification; meta-queries of the integration model; and unambiguous, formal semantics for the information integration. The current POPS data sources are the NASA X.500 Directory, the Competency Management System, and the Workforce Information Management System. Classes, properties, and relationships from three other databases will be joined into POPS over the next eighteen months.

Proactive Web – Goddard Space Flight Center

A "Proactive" web site is one that is capable of dynamically updating and modifying itself (content and structure). Thus, a proactive web site has the real potential to present to the user the most up-to-date information possible relating to the domains referenced by the web site. (This begins to directly address the problem statement at the beginning of this report.) The web site dynamics are realized through the unique integration and use of state-of-the-art semantic technologies mediated by a multi-agent system. The semantics of the web site will be initially captured in RDF and OWL. The agents utilize an adaptable user model with its ontologies (semantics) to drive the dynamics of the system. Since the user model is adaptable, then so are its associated ontologies. Thus, the project involves issues relating to dynamic ontology management. This activity is currently a collaboration between NASA/Goddard, the Pontifical University in Brazil, and the Worcester Polytechnic Institute. A prototype of such a web site is under development in order to demonstrate the idea and for initial performance evaluations. The prototype will be written in Java and should be available by the end of 2006.

SWEET – Earth Science, Jet Propulsion Laboratory

The Semantic Web for Earth and Environmental Terminology (SWEET) is an environment for sharing and organizing scientific knowledge. SWEET includes a collection of upper-level ontologies for Earth system science and related science data concepts represented in OWL to enable domain specialists to easily expand the contents. A collaboration Web site helps maintain alignment across registered ontologies enabling ontology updates to be propagated throughout the system where needed.

The NASA Taxonomy – Jet Propulsion Laboratory

Taxonomy development and Web information architecture provide a framework for Internet authors and service providers. Consistently modeled content makes it possible for engineers and scientists as well as the public to find and reuse content, rather than recreate it or make do without it. Adopting frameworks that enable content reuse provides an increased return on investment (ROI) for the time and effort spent to originally produce the material. It also allows content to be tagged for retrieval in larger contexts. A task to design a NASA Web Information Architecture was begun in FY 02. The scope of the task was modest and the primary deliverable at the end of the year was a Beta Taxonomy based on a series of interviews with NASA Subject Matter Experts from various communities. The vocabularies were designed to be broad enough to enable many integration points from disparate collections across the Agency. Aggregating materials from a number of sources and unifying them using a common reference model is key to NASA's ability to reuse solutions developed on one mission as they might apply to another. The NASA Taxonomy has a Core Metadata Specification and 11 facets with controlled vocabulary terms. The terms in the 11 core facets to the NASA Taxonomy are eventually meant to be applied to Agency wide ontologies, which will strengthen the semantic relationships, found across the organization.

Going Forward

Communities of practice are mostly aligned along project or programmatic lines. This isolates expertise and inhibits broader adoption of practical solutions. We should reach out these communities for a review of our analysis. A more cohesive community will create knowledge leverage, but is likely to be insufficient on it own given the size and rate of growth of our data problem. Additional human resources are required to effectively beat back the information tide. A mechanism to procure required talent and experience is needed.

We recommend that the NASA Enterprise Architect formally accept the direction articulated in this inaugural report. Moreover, we recommend the drafting of an overall vision of where we are going as an agency. There is often too much focus on failed strategies of the past or the pursuit of expensive solutions with limited benefit. Greater focus must be placed on practical phased pursuits that offer effective solutions and business benefits at each stage of implementation. This strategy should be vetted across NASA to give greater emphasis on practical pursuits that offer effective solutions. NASA should rely less on large-scale modifications to existing information services or grand consolidation efforts.

Next Steps: Short-term (18 months):

- Vet and review this approach with the KM community, NASA Earth Sciences Data Systems Working Group (DSWG), IEM and others.
- Brief the larger Information Resources community and seek official approval for approach.
- Join with the DSWG and others to form a community of practice.
- Define the Gold, Silver and Bronze criteria for NASA's Reference Model Types.
- Build a prototype repository service in collaboration with our community of practice.
- Assist developers in the construction of initial SLAPs for data and data model discovery and reuse.
- Assist developers in building a proof-of-concept repository for Ontologies and SLAPs and begin initial testing and requirements refinement.
- Learn how data validation is accomplished in flight hardware, flight software, science mission analysis, and others, and determine which best practices can be applied to the larger NASA data community.
- Construct go-to standards for new applications and models.
- Gain access to and participate in key W3C standards groups (e.g. WS-policy).
- Seek formal acceptance of our approach.

Follow-on Steps: 19 to 36 months:

- Create repository of ontologies, data reference models, and SLAPs.
- Develop strategies for adding metadata search and inference capabilities to designated document management, workflow management, and Exploration and science applications.
- Augment selected DRM models to include services like financial and risk calibrations, cross project schedule views, and Shuttle to CEV transitions.
- Refine the application architecture, identifying the initial set of candidate services to be deployed, and recommending the tools and standards to be used. Tools include those for ontology engineering and querying, development frameworks, inference engines, data stores, etc.
- Develop and deploy new applications using a Service-Oriented Architecture (SOA) approach; this will allow applications to access information from other applications in an ad hoc manner without having to retool and recode.
- Advertise applications and their interfaces using standards such as WSDL so they can be discovered automatically.

5-10 Years:

- Develop and deploy new classes of applications that merge data, services, and physical resources into a semantically aware, adaptive environment.
- Create a single pervasive collaborative environment by having software "tasking" agents autonomously scan published IT service assets (e.g., data repositories, projectors, displays, and printers) in conference areas, and choreograph them to an interconnected virtual work environment.
- Deploy software agents that can autonomously scan published knowledge and metadata and automatically connect them, or harvest them for information, anticipating users' needs: give the users the data they need when the need it, in a form relevant to their current task.
- Develop agents that can resolve conflicts amongst different data sources and ascertain the trustworthiness of the published data, both within NASA and outside the Agency.
- Develop agents that can learn, anticipate needs, discover relevant data, and enter into transactions, all on behalf of their human users.

GLOSSARY

COI – Community of Interest. A group with a common interest.

COP – Community of Practice. A group with a common applied interest.

RDF – Resource Description Framework. A general-purpose language for representing information in the Web.

RIF – Rules Interchange Format. The W3C released their first draft of this standard in March of 2006.

OWL – Web Ontology Language. A W3C recommendation from 2004. The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S), by providing additional vocabulary along with formal semantics. *

SPARQL - (recursively, SPARQL Protocol and RDF Query Language) is a Semantic Web candidate recommendation presently (as of 2006) undergoing standardization by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium.*

SWRL – Semantic Web Rule Language. Based on a combination of the OWL DL and OWL Lite sublanguages of the OWL Web Ontology Language with the Unary/Binary Datalog RuleML sublanguages of the Rule Markup Language. (from W3C 2004).

Acknowledgments

Some definitions for the glossary and elsewhere came from wikipedia.

EA Data Team Members
Les Farkas, GRC
Celeste Merryman, ARC
Mark Page, KSC
Andrew Schain, HQ
Bill Seixas, JPL

Appreciation to our editors and technical advisors
Kendall Clark
Bijan Parsia
Doug Ward
Rob Winters

Data Model Contributers
Special thanks to those that contributed write ups of their data models (without reading the paper in advance!). Their contributions reflect a willingness to share but do not equate to full agreement regarding the content of this report.

Bruce Altner, HQ
Jayne Dutra, JPL
Jonathan Pettus, MSFC
Alex Pline, HQ
Rob Raskin, JPL
Bob Smith, HQ
Walt Truszkowski, GSFC
Brian Wilson, JPL